GENE

[Notice-MY-2023-03; Docket No. 2023-0002; Sequence No.37]

Office of Shared Solutions and Performance Improvement (OSSPI); Chief Data Officers Council (CDO); Request for Information - Synthetic Data Generation

AGENCY: Federal Chief Data Officers (CDO) Council; General Services Administration, (GSA).

ACTION: Notice.

SUMMARY: The Federal CDO Council was established by the Foundations for Evidence-Based Policymaking Act. The Council's vision is to improve government mission achievement and increase benefits to the nation through improving the management, use, protection, dissemination, and generation of data in government decision-making and operations. The CDO Council is publishing this Request for Information (RFI) for the public to provide input on key questions concerning synthetic data generation. Responses to this RFI will inform the CDO Council's work to establish best practices for synthetic data generation.

DATES: We will consider comments received by [INSERT DATE 30 DAYS AFTER DATE OF PUBLICATION IN THE **FEDERAL REGISTER**].

Targeted Audience

This RFI is intended for Chief Data Officers, data scientists, technologists, data stewards and data-and evidence-building related subject matter experts from the public, private, and academic sectors.

ADDRESSES: Respondents should submit comments identified by Notice-MY-2023-03 via the Federal eRulemaking Portal at *https://www.regulations.gov* and follow the instructions for submitting comments. All public comments received are subject to the Freedom of Information Act and will be posted in their entirety at *regulations.gov*, including any personal and/or business confidential information provided. Do not include any information you would not like to be made publicly available.

Written responses should not exceed six pages, inclusive of a one-page cover page as described below. Please respond concisely, in plain language, and specify which question(s) you are responding to. You may also include links to online materials or interactive presentations, but please ensure all links are publicly available. Each response should include:

- The name of the individual(s) and/or organization responding.
- A brief description of the responding individual(s) or organization's mission and/or areas of expertise.
- The section(s) that your submission and materials are related to.
- A contact for questions or other follow-up on your response.

By responding to the RFI, each participant (individual, team, or legal entity) warrants that they are the sole

author or owner of, or has the right to use, any copyrightable works that the submission comprises, that the works are wholly original (or is an improved version of an existing work that the participant has sufficient rights to use and improve), and that the submission does not infringe any copyright or any other rights of any third party of which participant is aware.

By responding to the RFI, each participant (individual, team, or legal entity) consents to the contents of their submission being made available to all Federal agencies and their employees on an internal-to-government website accessible only to agency staff persons.

Participants will not be required to transfer their intellectual property rights to the CDO Council, but participants must grant to the Federal Government a nonexclusive license to apply, share, and use the materials that are included in the submission. To participate in the RFI, each participant must warrant that there are no legal obstacles to providing the above-referenced nonexclusive licenses of participant rights to the Federal Government. Interested parties who respond to this RFI may be contacted for follow-on questions or discussion.

FOR FURTHER INFORMATION CONTACT: Issues regarding submission or questions can be sent to Ken Ambrose and Ashley Jackson, Senior Advisors, Office of Shared Solutions and Performance Improvement, General Services

Administration, at 202-215-7330 (Kenneth Ambrose) and 202-538-2897(Ashley Jackson), or *cdocstaff@gsa.gov*.

SUPPLEMENTARY INFORMATION:

Background

Pursuant to the Foundations for Evidence-Based Policy Making Act of 2018,[1] the CDO Council is charged with establishing best practices for the use, protection, dissemination, and generation of data in the Federal Government. In reviewing existing activities and literature from across the Federal Government, the CDO Council has determined that:

- the Federal Government would benefit from developing consensus of a more formalized definition for synthetic data generation,
- synthetic data generation has wide-ranging applications, and
- there are challenges and limitations with synthetic data generation.

The CDO council is interested in consolidating feedback and inputs from qualified experts to gain additional insight and assist with establishing a best practice guide around synthetic data generation. The CDO Council has preliminarily drafted a working definition of synthetic data generation and several key questions to better inform its work.

[1] H.R.4174 - 115th Congress (2017-2018): Foundations for Evidence-Based Policymaking Act of 2018 | Congress.gov | Library of Congress https://www.congress.gov/bill/115th-congress/house-bill/4174/text

Information and key questions

Section 1: Defining synthetic data generation

Synthetic data generation is an important part of modern data science work. In the broadest sense, synthetic data generation involves the creation of a new synthetic or artificial dataset using computational methods. Synthetic data generation can be contrasted with real-world data collection. Real-world data collection involves gathering data from a first-hand source, such as through surveys, observations, interviews, forms, and other methods. Synthetic data generation is a broad field that employs varied techniques and can be applied to many different kinds of problems. Data may be fully or partially synthetic. A fully synthetic dataset wholly consists of points created using computational methods, whereas a partially synthetic dataset may involve a mix of first-hand and computationally generated synthetic data.

Throughout this RFI, we use the following definitions:

- data - recorded information, regardless of form or the media on which the data is recorded[2]
- data asset - a collection of data elements or data sets that may be grouped together[3]
- open government data asset - a public data asset that is (A)machine-readable; (B)available (or could be made available) in an open format; (C)not encumbered by

[2] 44 U.S.C. 3502(16)
[3] 44 U.S.C. 3502(17)

restrictions, other than intellectual property rights, including under titles 17 and 35, that would impede the use or reuse of such asset; and (D)based on an underlying open standard that is maintained by a standards organization[4]

The National Institute of Standards and Technology (NIST) defines synthetic data generation as "a process in which seed data is used to create artificial data that has some of the statistical characteristics as the seed data".[5]

The CDO Council believes that this definition of synthetic data generation includes techniques such as using statistics to create data from a known distribution, generative adversarial networks (GANs),[6] variational autoencoding (VAE),[7] building test data for use in software development,[8] privacy-preserving synthetic data generation[9] and others.

The CDO Council also believes that it is important to draw contrasts between synthetic data generation and other activities. For example, synthetic data generation does not include collection of data without any inference. Synthetic data generation does not include signal processing, such as automated differential translations of global positioning

---

[4] 44 U.S.C. 3502(20)
[5] https://csrc.nist.gov/glossary/term/synthetic_data_generation
[6] 15 U.S.C. 9204
[7] A useful definition of this technique is available in the abstract of this paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8774760/
[8] This technique is described in the Department of Defense DevSecOps Fundamentals Guidebook https://dodcio.defense.gov/Portals/0/Documents/Library/DevSecOpsTools-ActivitiesGuidebook.pdf, page 23
[9] NIST Special Publication 800-188, Section 4.4 https://doi.org/10.6028/NIST.SP.800-188

satellite data. Synthetic data generation also does not include enriching data during data analysis – such intermediate steps that involve augmenting or enhancing existing data but do not involve the creation of artificial data.

Other analysis techniques, such as distribution fitting and parametric modeling, are closely related to synthetic data generation. The CDO Council believes the key difference; however, is the purpose of the computational methods. Synthetic data generation seeks to create wholly new data points based on the statistical properties of a dataset, whereas distribution fitting seeks to 'fill in' a dataset based on a known distribution. Notably, the fitted distribution can be used to generate points that are not part of the original dataset – which is an application of synthetic data generation.

Questions:

- Are there any limitations to relying on the NIST definition to describe the field of synthetic data generation? How should it be improved?
- How well does the CDO Council's list of examples and contrasts improve understanding? How should these be improved?

Section 2: Applying synthetic data generation

Synthetic data generation can enable the creation of larger and more diverse datasets, enhance model performance, and

protect individual privacy. The CDO Council's review of potential applications of synthetic data generation found examples in:

- Data augmentation.[10] This application involves creating new data points or datasets from existing data. This application can be particularly useful in developing training datasets for machine learning and advanced analytics.
- Data synthesis.[11] This application involves using an existing dataset to create a new dataset, sharing similar statistical properties with the original dataset, to protect individual privacy. Generating such datasets has wide-ranging applications including, but not limited to, facilitating reproducible investigation of clinical data while preserving individual privacy.
- Modeling and simulation.[12] This application involves setting assumptions, parameters and rules to develop data for further analysis. The synthetic dataset can be used for developing insights, testing hypotheses, and/or understanding a model's behavior. This application supports the conduct of controlled experiments, predicting potential future outcomes from

[10] This application is briefly described at https://frederick.cancer.gov/initiatives/scientific-standards-hub/ai-and-data-science, Section 4
[11] A definition of this technique is available in the abstract of this paper https://par.nsf.gov/servlets/purl/10187206
[12] A definition a computer simulation is proposed at https://builtin.com/hardware/computer-simulation

current conditions, generating scenarios for rare or extreme events, and validating or calibrating a model.

- Software development.[13] This application involves using existing database schemas to simulate real-world scenarios and ensure that a software application can handle different types of data and errors effectively. This application assists in the creation of representative data, makes it easier to generate edge cases, protects individual privacy, and improves testing efficiency.

Notably, the CDO Council believes that not all applications of modeling and simulation would meet the definition of synthetic data generation. For example, weather forecasting applies numerical models and applies a complex mix of data analysis, meteorological science, and computation methods but does not involve the creation of synthetic or artificial data points. Instead, the purpose of these models is to predict future conditions.

Questions:

- How are these examples representative of synthetic data generation? How should they be revised?
- What other examples of synthetic data generation should the CDO Council know about?
- What are the key advantages for the use of synthetic data generation?

[13] DoD DevSecOps Fundamentals, ibid.

Section 3: Challenges and limitations in synthetic data generation

The CDO Council recognizes that synthetic data generation can be a valuable technique. However, it should be noted that there are some challenges and limitations with the technique. For example, there can be challenges generating data that realistically simulates the real world and the diversity of real data. Additionally, evaluating the quality of a synthetic dataset may also be extremely challenging.

Synthetic data generation is also subject to challenges commonly facing any statistical methods, such as overfitting and imbalances in the source data. These challenges reduce the utility of the generated synthetic data because they may not be properly representative, including failing to represent rare classes.

Questions:

- What other challenges and limitations are important to consider in synthetic data generation?
- What tools or techniques are available for effectively communicating the limitations of generated synthetic data?
- What are best practices for CDOs to coordinate with statistical officials on synthetic data?
- What approaches can CDOs consider to help address these challenges?

Section 4: Ethics and equity considerations in synthetic data generation

Synthetic data generation techniques hold great promise, but also introduce questions of ethics and equity. Consistent with Federal privacy practices,[14] any data generation technique involving individuals must respect their privacy rights and obtain informed consent before using real-world data to generate synthetic data. As noted in Section 3, synthetic data generation is also subject to challenges commonly facing any statistical methods and has the potential to introduce and encode errors or bias, potentially leading to discriminatory outcomes.

Uses of generated synthetic data must also be carefully considered. The context and quality of the generated synthetic data will impact its practical utility and impact. Assessing and understanding the fitness of a generated synthetic dataset is essential. For instance, a generated synthetic dataset may not sufficiently represent the diversity of the source dataset. In addition, a generated synthetic dataset may not contain sufficient variables to fully represent the system and the drivers of differences in the phenomenon it is meant to represent.

Questions:

- What techniques are available to facilitate transparency around generated synthetic data?

[14] OMB Circular A-130, Appendix II https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf

- What are best practices for CDOs to coordinate with privacy officials on ethics and equity matters related to synthetic data generation?
- How can we apply the Federal Data Ethics Framework[15] to address these ethics and equity concerns?

Section 5: Synthetic data generation and evidence-building

Synthetic data generation can enable the production of evidence for use in policymaking. Applications such as simulation or modeling can help policymakers explore scenarios and their potential impacts. Likewise, policymakers can conduct controlled experiments of potential policy interventions to better understand their impacts. Data synthesis may help policymakers make more data publicly available to spur research and other foundational fact-finding activities that can inform policymaking.

Questions:

- What other applications of synthetic data generation support evidence-based policymaking?[16]
- What is the relationship between synthetic data generation and open government data?[17]
- How can CDOs and Evaluation Officers best collaborate on synthetic data generation to support evidence-

[15] https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf
[16] OMB Memorandum M-19-23
[17] 44 U.S.C. 3520(20)

building?[18] What about other evidence officials?[19]

Kenneth Ambrose,

Senior Advisor CDO Council,

Office of Shared Solutions and Performance Improvement,

General Services Administration.

Billing Code: 6820-69

[FR Doc. 2024-00036 Filed: 1/4/2024 8:45 am; Publication Date: 1/5/2024]

[18] OMB Memorandum M-19-23, Appendix A
[19] OMB Memorandum M-19-23, Section II (Key Senior Officials)